Inter-Task Association Critic for Cross-Resolution Person Re-Identification

Zhiyi Cheng
Queen Mary University of London
z.cheng@qmul.ac.uk

Qi Dong
Queen Mary University of London
q.dong@qmul.ac.uk

Shaogang Gong
Queen Mary University of London
s.gong@qmul.ac.uk

Xiatian Zhu
Vision Semantics Limited, London, UK
eddy.zhuxt@gmail.com

Abstract

Person images captured by unconstrained surveillance cameras often have low resolutions (LR). This causes the resolution mismatch problem when matched against the high-resolution (HR) gallery images, negatively affecting the performance of person re-identification (re-id). An effective approach is to leverage image super-resolution (SR) along with person re-id in a joint learning manner. However, this scheme is limited due to dramatically more difficult gradients backpropagation during training. In this paper, we introduce a novel model training regularisation method, called Inter-Task Association Critic *(INTACT), to address this fundamental problem. Specifically, INTACT discovers the underlying association knowledge between image SR and person re-id, and leverages it as an extra learning constraint for enhancing the compatibility of SR model with person re-id in HR image space. This is realised by parameterising the association constraint which enables it to be automatically learned from the training data. Extensive experiments validate the superiority of INTACT over the state-of-the-art approaches on the cross-resolution re-id task using five standard person re-id datasets.*

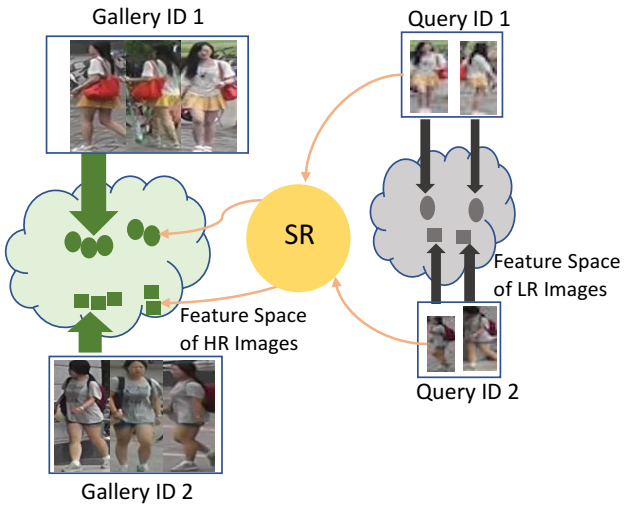

Figure 1. Illustration of cross-resolution person re-identification. Resolution mismatch between the low-resolution (LR) query images and the high-resolution (HR) gallery images causes unaligned feature distributions and finally inferior identity matching performance. One effective solution is using an image super-resolution (SR) model to enhance the resolution of LR query images for alleviating the distribution discrepancy with HR gallery images.

1. Introduction

Person re-identification (re-id) aims to match the identity information in the images captured by disjoint surveillance camera views [13]. Most existing methods assume that the probe and gallery images have similar and sufficiently high resolutions. However, due to unconstrained distances between cameras and pedestrians, person images are often captured at various resolutions. This *resolution mismatch* issue brings about significant challenges to re-id. As low-resolution (LR) images contain much less identity detail information than high-resolution (HR) images, directly matching them across resolutions leads to substantial performance drop [16, 26]. For example, a standard person re-id model [12] can suffer up to 19.2% Rank-1 rate drop when applied to cross-resolution person re-id [26].

A number of cross-resolution re-id methods have been developed for addressing the resolution mismatch problem [7, 16, 26, 40]. They are generally in two categories: (1) Learning resolution-invariant representation [7] and (2) Exploiting image super-resolution (SR) [16, 40]. The first category aims at learning a feature representation space shared by LR and HR images, but tends to lose fine-grained discriminative details due to being absent in LR images. The second category can solve this limitation often by adopting a multi-task joint learning framework which cascades

SR and re-id. However, this design suffers from ineffective model training due to significantly higher difficulty of backpropagating the gradients through such a cascaded thus heavier model [2]. As a consequence, the SR model is less compatible with person re-id. Recently, Li et al. [26] combined the two approaches in a unified framework for improving cross-resolution re-id performance, but still leaving the above problem unsolved.

In this work, we address this problem by introducing a novel regularisation named *Inter-Task Association Critic* (INTACT). It is an inter-task association mechanism that smooths out two unique tasks in joint learning. In design, it consists of a cascaded multi-task (SR & re-id) network and an association critic network. The objective is to enhance the compatibility between SR and re-id, *i.e.*, super-resolving LR person images in such a way that the resolved images are suited for the re-id model to perform identity matching in the HR image space. This is realised in two parts by INTACT: **(I)** We parameterise the (unknown) inter-task association constraint with a dedicated network, which enables it to be learned directly from the HR training data. **(II)** Once learned, serving in a critic role the association constraint is then applied to supervise the SR model. That means, the SR model training is further constrained to satisfy the learned inter-task association.

We make three **contributions**: **(1)** We propose an idea of leveraging the association between image SR and person re-id tasks for solving the under-studied yet significant cross-resolution re-id problem. **(2)** We formulate a novel regularisation method, called *Inter-Task Association Critic* (INTACT), for implementing the proposed inter-task association. It is established via parameterising the association, and end-to-end trainable. **(3)** Extensive experiments show the performance advantages of our INTACT over a wide range of state-of-the-art methods on five person re-id benchmarks in the cross-resolution person re-id problem.

2. Related Work

Person re-id There are an increasing number of studies for person re-identification in the past decade [1, 22, 26, 35, 36, 43, 48, 49, 8, 3]. Many of the existing works focus on addressing re-id challenges from variations in background clutter [23], human poses [28], or occlusion [32] across camera views. More advanced network structures [52, 8, 24] have been developed to boost the matching accuracy. Besides, there are extensive efforts on unsupervised learning [22, 21, 42, 5, 30, 27], domain adaptation [50, 6, 37, 33, 45, 41, 46], weak supervision [54, 31] for minimising the labelling efforts, and text-image person search [11, 44]. Among these, a largely ignored aspect in re-id is that, persons images from unconstrained surveillance cameras often have varying resolutions, which would degrade the model performance if not properly handled.

Cross-resolution person re-id To address the resolution mismatch problem, several cross-resolution person re-id methods have been proposed [7, 16, 25, 26, 40]. They are fallen into two groups: (1) Learning resolution-invariant representation [7, 17, 25] and (2) Exploiting image super-resolution (SR) [16, 40]. In the *first* group, Jing et al. [17] propose to learn the mapping between HR and LR representations by a semi-coupled low-rank dictionary learning model; Li et al. [25] align the cross-resolution representation with a heterogeneous class mean discrepancy criterion. Chen et al. [7] learn the resolution-invariant representation by adding an adversarial loss on the representation features of HR and LR images. A weakness of these methods is that, such learned representations involve only coarse appearance information sub-optimal for re-id. That is because fine-grained details, lacking in LR images but rich in HR images, are thrown away during learning for an agreement.

The *second* group of models, designed to exploit image super-resolution, can solve this limitation. Both methods [16, 40] adopt a joint learning strategy of SR and re-id in a cascade, integrating identity-matching constraints with SR learning end-to-end. However, this design suffers from ineffective model training due to significantly higher difficulty of back-propagating the gradients through such a cascaded heavy model [2, 20]. Recently, Li et al. [26] combine the resolution-invariant representations with those exacted from resolution-recovered images, and achieve state-of-the-art performance. However, the above problem remains unsolved. To that end, in this work we introduce a novel regularisation based on an inter-task association mechanism.

Image super-resolution and recognition The low-resolution object recognition problem has drawn attentions in recent years [4, 9, 29, 34, 38, 47]. Broadly, there are other studies [9, 47, 10] that unite image SR and object recognition in a multi-task joint-learning framework. Nonetheless, they share the same learning limitation as [16, 40]. Therefore, the proposed method is potentially beneficial to these works conceptually.

3. Methodology

Problem setting We consider the cross-resolution person re-id problem. In model training, we assume a set of identity labelled high-resolution (HR) training images $\mathcal{D} = \{x_h, y\}$. The *objective* is to learn a person re-id model that can tackle low-resolution (LR) query images in matching against a set of HR gallery images at test time.

We explore the potential of image super-resolution (SR). The intuition is that an effective SR model should be able to recover the resolution of LR images so that the resolution mismatch problem between the query and gallery images can be well alleviated. To encourage that the SR model can

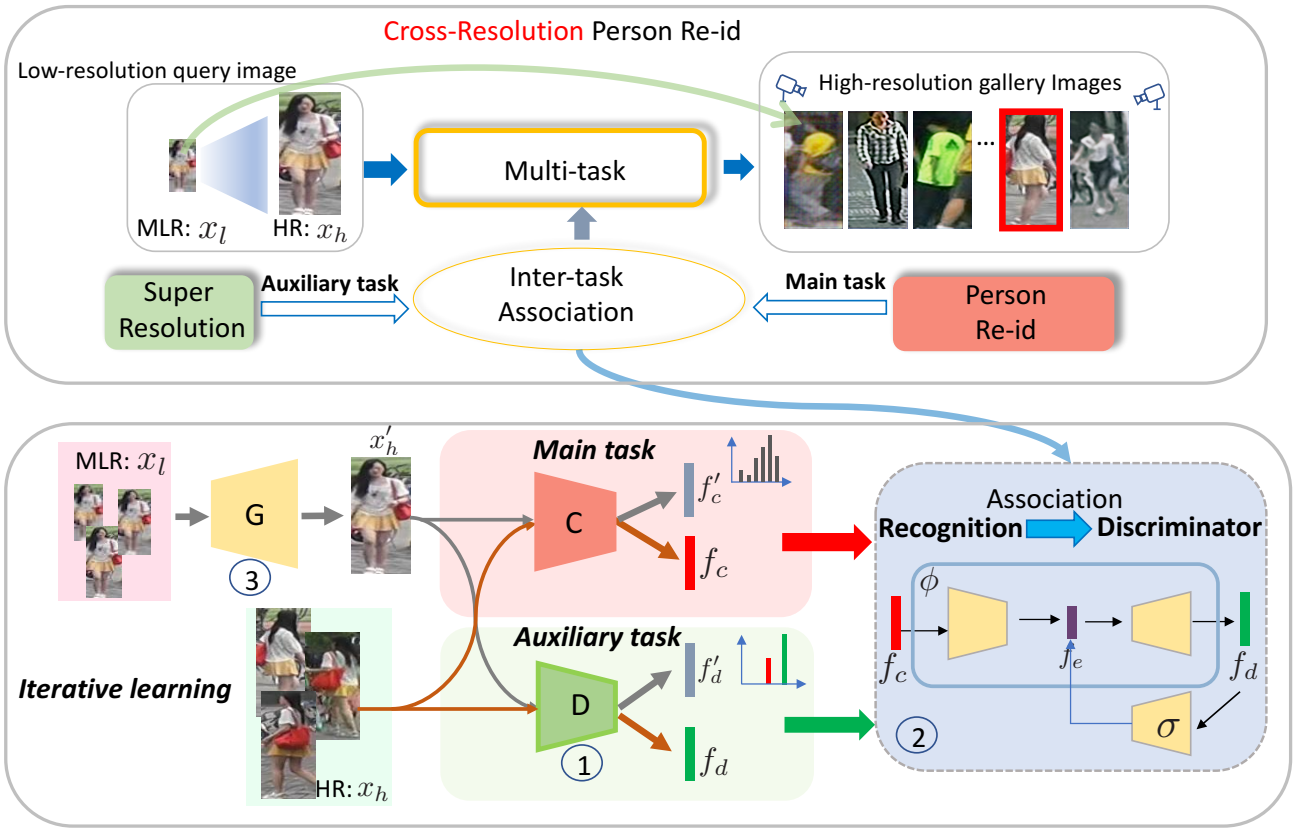

Figure 2. Overview of the proposed *Inter-Task Association Critic* (INTACT) method for cross-resolution person re-id. Specifically, INTACT aims to recover the resolution of LR query images in such a way that the super-resolved images can be more accurately matched against HR gallery images for person re-id. Joint learning of an image super-resolution (SR) model and a person re-id model in a cascaded manner is *unsatisfactory*, due to higher difficulty of backpropagating the gradients through two cascaded models. INTACT offers a superior solution. Our model is trained alternatively in three steps: **(1)** Update the discriminator D of a GAN model; **(2)** Update the inter-task association module ϕ between the identity recognition representation $\boldsymbol{f}_{\rm c}$ and discriminator representation $\boldsymbol{f}_{\rm d}$. **(3)** Update the generator G of the GAN model, subject to the learned association regularisation on the identity recognition representation $\boldsymbol{f}'_{\rm c}$ and discriminator representation $\boldsymbol{f}'_{\rm d}$ of the resolved images. G: the generator of GAN model; D: the discriminator; C: the person re-id model trained with cross-entropy loss.

generate such HR images that are more effective for person re-id, a straightforward approach is to form a joint multi-task learning pipeline by cascading SR and re-id sequentially, as exemplified in [16].

3.1. Joint Multi-Task Learning

Image super-resolution model To train a SR model, we typically use a set of LR-HR image pairs $\{(x_l, x_h)\}$ with pixel alignment. Often, we form such pairs by downsampling the HR training images. In this study, we choose the Generative Adversarial Network (GAN) model [14] for SR due to its promising performance [19].

GAN solves a min-max optimisation problem, where the discriminator D aims to distinguish the real HR from super-resolved images, while the generator G aims for generating super-resolved images that can fool the discriminator. The objective function can be defined as:

$$\mathcal{L}_{\rm gan} = \mathbb{E}_{x_h}[\log D(x_h)] + \mathbb{E}_{x_l}[\log\left(1 - D(G(x_l))\right] \quad (1)$$

More specifically, the generator G tries to minimise the objective value against an adversarial discriminator D that instead tries to maximise the value. The optimal solution is obtained as:

$$G^* = \arg\min_G \max_D \mathcal{L}_{\rm gan}. \quad (2)$$

Person re-id model With the training data $\mathcal{D}$, one can train any existing person re-id model (*e.g.* [52]) by a softmax Cross-Entropy loss function:

$$\mathcal{L}_{\rm id} = -\log(p_y), \quad (3)$$

where y is the ground-truth person identity of x_l, and p_y the prediction probability on class y.

Joint multi-task learning To build a joint multi-task learning pipeline, we can simply cascade SR and re-id by using the output $G(x_l)$ of the SR as the input of re-id model. The overall objective function is then formulated as:

$$\mathcal{L}_{\text{sr}} = \mathcal{L}_{\text{MSE}} + \lambda_g \mathcal{L}_{\text{gan}} + \lambda_c \mathcal{L}_{\text{id}} \tag{4}$$

where $\mathcal{L}_{\text{MSE}}$ is the pixel-wise content loss, defined as $\mathcal{L}_{\text{MSE}} = \|x_h - G(x_l)\|_2^2$. λ_g and λ_c are weight parameters.

Limitation Despite a good solution for cross-resolution re-id, this pipeline is intrinsically limited. This is due to significantly higher difficulty of backpropagating the gradients through two cascaded models [2, 20]. As a consequence, the SR model training is not properly constrained for maximising the person re-id performance, *i.e*. the resulted SR model is not well compatible with the re-id model.

3.2. Inter-Task Association Critic

To address this fundamental limitation, we introduce a novel regularisation, *Inter-Task Association Critic* (INTACT). The key idea is to exploit the intrinsic association between the SR and re-id tasks as an extra optimisation constraint for boosting their joint learning and enhancing their compatibility. However, it is nontrivial to quantify such inter-task association which are typically complex and unknown a priori. To solve this issue, we propose to parameterise this association.

Specifically, we leverage a dedicated network to represent the association from the main task (*i.e*. person re-id) to the auxiliary task (*i.e*. SR). During model training, INTACT consists of two parts. In part **I**, it discovers the association using the native HR images. Concretely, it learns the association network residing between the discriminator and identity classification representations on the HR training images $\{x_h\}$. In part **II**, the learned association is then applied as a regularisation in the SR model training. Concretely, the discriminator and classification representations extracted from the resolved images are encouraged to satisfy the association constraint pre-learned from the true HR data. An overview of our INTACT is depicted in Fig. 2.

Part I: Association Learning With a GAN model we represent the real-fake judgement task by the feature activation $\boldsymbol{f}_{\text{d}}$ of the discriminator. For the identity classification task, a large number of on-the-shelf person re-id models can be adopted. We exploit a very recent method presented in [52] for extracting identity classification feature $\boldsymbol{f}_{\text{c}}$ to represent identity. We train the re-id model using HR images x_h independently to achieve the best identity representing power. It is trained one-off, frozen and served as an identity critic for the following model optimisation.

Given an input LR image x_l, we want the generator (SR model) to output a super-resolved HR image $G(x_l)$ with high identity discrimination. To achieve this, we propose to design an association constraint ϕ between the real-fake discriminator representation $\boldsymbol{f}_{\text{d}}$ and identity classification $\boldsymbol{f}_{\text{c}}$ representations of the image x_h. Then, we represent and learn ϕ on HR training images x_h with a small network, considering that they are the target the SR images $G(x_l)$ need to approach during training.

Formally, we learn the association as the transformation from the identity recognition $\boldsymbol{f}_{\text{c}}$ to discriminator $\boldsymbol{f}_{\text{d}}$ representations. This is based on a hypothesis that the identity recognition representation, learned from HR training images, contains the information for general high-resolution distribution (that the real-fake discriminator tries to learn); Whilst the discriminator features are relatively less informative compared to the identity ones, due to being derived from a simpler binary classification task (real or fake). Learning such a mapping is thus more sensible. In particular, we derive an association regularisation as:

$$\mathcal{L}_{\text{intact}} = ||\phi(\boldsymbol{f}_{\text{c}}) - \boldsymbol{f}_{\text{d}}||_2^2 \tag{5}$$

It aims to optimise the parameters ϕ of the association network, using $\boldsymbol{f}_{\text{c}}$ extracted from the re-id model and $\boldsymbol{f}_{\text{d}}$ extracted from the discriminator on the HR image x_h.

To facilitate learning ϕ, an additional bridging constraint is further imposed for manipulating the optimising direction. Specifically, we isolate an intermediate latent feature space $\boldsymbol{f}_{\text{e}}$ from ϕ such that a bridging operation can be implanted with a transform σ of the target $\boldsymbol{f}_{\text{d}}$, defined as:

$$\mathcal{L}_{\text{e}} = ||\sigma(\boldsymbol{f}_{\text{d}}) - \boldsymbol{f}_{\text{e}}||_2^2 \tag{6}$$

where $\boldsymbol{f}_{\text{e}}$ is obtained in the middle latent space of ϕ with $\boldsymbol{f}_{\text{c}}$ as input. The bridging module σ is jointly learned with the association module ϕ in a combination as:

$$\mathcal{L}_{\text{intact-e}} = \mathcal{L}_{\text{intact}} + \mathcal{L}_{\text{e}} \tag{7}$$

Part II: Association Regularisation Once the inter-task association network ϕ is learned as above, we treat it as a critic to regularise the learning of the SR model (the generator) and the discriminator in the GAN based multi-task learning network. We distil the learned association by similarly coupling the information of discriminator and identity recognition. Particularly, this distillation loss is in the same form of Eq. (5) but applied to the SR images $G(x_l)$ as:

$$\mathcal{L}_{\text{dis}} = ||\phi(\boldsymbol{f}'_{\text{c}}) - \boldsymbol{f}'_{\text{d}}||_2^2 \tag{8}$$

where $\boldsymbol{f}'_{\text{c}}$ and $\boldsymbol{f}'_{\text{d}}$ are the corresponding identity and discriminator representations of a single SR image $G(x_l)$ analogue to x_h above.

It is worth mentioning that, unlike Eq. (5) here we frozen the association network ϕ that functionally serves as an external *critic* in this step. This role is similar in spirit as the ImageNet pretrained VGG model of the perceptual loss

[18]. Using $\mathcal{L}_{\text{dis}}$ along with GAN training, we essentially encourage the synthesis of such HR images that respect the same association relation between identity and fidelity on the genuine HR images. This is the key drive behind our INTACT model that imposes both supervision signals and importantly their interaction in a single formulation.

Remarks Unlike the *de facto* multi-task inference using weighted loss summation for inter-task interaction learning and communicating, we discover the underlying association between two tasks as an extra learning constraint. Significantly, once parameterised this association can be automatically learned from the original training data themselves in a data-driven manner, without any hand-crafting and the need for ad-hoc knowledge. Consequently, the intrinsic conflicts between two different tasks can be mitigated effectively, benefiting the overall model learning process towards person identity matching. Moreover, we can also consider that INTACT takes a *soft* integration design that aims to link the underlying objectives between two different tasks by maximising their positive correlation during training. Consequently, the two learning objectives can adaptively collaborate in a unified learning process with a balanced trade-off between individual and common pursuits.

3.3. Model Training

In model training, our INTACT loss terms are seamlessly integrated with the standard GAN optimisation with one more step. The whole model remains end-to-end trainable. The entire training process is summarised in Algorithm 1.

Algorithm 1 INTACT model training

Input: Training data $\mathcal{D} = \{x_l, x_h\}$ with identity labels Y.
Output: A person image super-resolution (SR) model.
Initialisation: Training a standard person re-id model with HR images and the identity labels.
Alternating training (frozen one, and update the others):
for $i = 1$ **to** *iter* **do**
(1) Update the discriminator with the GAN loss (Eq. (2));
(2) Update the association network ϕ (Eq. (7));
(3) Update the generator (SR model) with the SR objective loss (Eq. (4)) and distillation loss (Eq. (8)).
end for

4. Experiments

4.1. Datasets

We used five person re-id benchmarks in our evaluations. The **CUHK03** dataset comprises 14,097 images of 1,467 identities with 5 different camera views. As [26], we used the 1,367/100 training/test identity split. The **VIPeR** dataset contains 632 person image pairs captured by 2 cameras. Following [26], we randomly divided this dataset into two non-overlapping halves based on the identity labels. Namely, images of a subject belong to either the training or the test set. The **CAVIAR** dataset contains 1,220 images of 72 person identities captured by 2 cameras. We discarded 22 people who only appear in the closer camera, and split the remaining into two non-overlapping halves in the identity labels as [26]. The **Market-1501** dataset consists of 32,668 images of 1,501 identities captured in 6 camera views. We used the standard 751/750 training/test identity split. The **DukeMTMC-reID** dataset contains 36,411 images of 1,404 identities captured by 8 cameras. We adopted the standard 702/702 training/test identity split.

Following [16, 26], we evaluated the setting of multiple low-resolution (MLR) person re-id. We tested four synthetic and one real-world cross-resolution re-id benchmarks. Specifically, for the synthetic cases (Market-1501, CUHK08, VIPeR, and DukeMTMC), the query images taken from one camera are down-sampled by a randomly selected downsampling rate r $\in \{2, 3, 4\}$ (*i.e.* the spatial size of a downsampled image becomes H/r $\times$ W/r), while the images taken by the other camera(s) remain unchanged. We name the Multiple Low Resolution (MLR) datasets as ***MLR-dataset***. On the other hand, the CAVIAR dataset provides realistic images of multiple resolutions, *i.e.* a genuine MLR dataset for evaluating cross-resolution person re-id.

4.2. Experimental Settings

We evaluated the proposed INTACT model using the cross-resolution person re-id setting [16, 26], where the probe set contains LR images whilst the gallery set contains HR images. We adopted the standard single-shot person re-id setting, and used the average cumulative match characteristic as the evaluation metric.

4.3. Implementation Details

We performed all the experiments in PyTorch on a machine with a Tesla P100 GPU. During training, the varying LR images are generated by randomly down-sampling HR images by $r \in \{2, 3, 4\}$ times. All the LR images were then resized to 256×128×3 for both model training and deployment. We used the residual blocks [15] as the backbone of our model. For the SR generator, we adopted an encoder-decoder architecture. Specifically, it consists of 16 residual blocks equally distributed in 8 groups. The resolution drops 16 times from 256×128 to 16×8 pixels (due to the first 4 residual block groups each with a max pooling layer), and then increases back to 256×128 with the last 4 groups of residual block each with pixel shuffling. The generator's architecture is shown in Fig. 3. The discriminator is similar as [19]. The person re-id network [52] was trained using HR data. Once trained, it was frozen when training INTACT.

We implemented the inter-task association network ϕ by

Table 1. Cross-resolution person re-id performance (%). Bold and underlined numbers indicate top two results, respectively.

Model	MLR-Market-1501			MLR-CUHK03			MLR-VIPeR			MLR-DukeMTMC-reID			CAVIAR		
	Rank1	Rank5	Rank10	Rank1	Rank5	Rank10	Rank1	Rank5	Rank10	Rank1	Rank5	Rank10	Rank1	Rank5	Rank10
CamStyle [51]	74.5	88.6	93.0	69.1	89.6	93.9	34.4	56.8	66.6	64.0	78.1	84.4	32.1	72.3	85.9
FD-GAN [12]	79.6	91.6	93.5	73.4	93.8	97.9	39.1	62.1	72.5	67.5	82.0	85.3	33.5	71.4	86.5
SLD^{2}L [17]	-	-	-	-	-	-	20.3	44.0	62.0	-	-	-	18.4	44.8	61.2
SING [16]	74.4	87.8	91.6	67.7	90.7	94.7	33.5	57.0	66.5	65.2	80.1	84.8	33.5	72.7	89.0
CSR-GAN [40]	76.4	88.5	91.9	71.3	92.1	97.4	37.2	62.3	71.6	67.6	81.4	85.1	34.7	72.5	87.4
JUDEA [25]	-	-	-	26.2	58.0	73.4	26.0	55.1	69.2	-	-	-	22.0	60.1	80.8
SDF [39]	-	-	-	22.2	48.0	64.0	9.3	38.1	52.4	-	-	-	14.3	37.5	62.5
RAIN [7]	-	-	-	78.9	97.3	98.7	42.5	68.3	79.6	-	-	-	42.0	77.3	89.6
CAD [26]	83.7	92.7	95.8	82.1	97.4	**98.8**	43.1	68.2	77.5	75.6	86.7	89.6	42.8	76.2	91.5
INTACT (Ours)	**88.1**	**95.0**	**96.9**	**86.4**	**97.4**	98.5	**46.2**	**73.1**	**81.6**	**81.2**	**90.1**	**92.8**	**44.0**	**81.8**	**93.9**

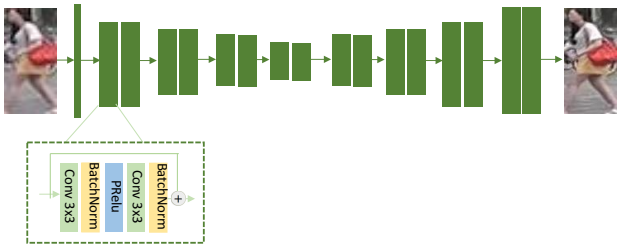

Figure 3. Architecture of image SR model (GAN's generator).

an encoder-decoder network, where the intermediate latent feature space $\boldsymbol{f}_{\mathrm{e}}$ is set as the encoder's output. Both encoder and decoder contain three FC layers each followed with batch-normalisation, respectively. The dimension of the latent feature space was set to 200. The bridging module σ for $\boldsymbol{f}_{\mathrm{d}}$ shares the same structure to the encoder of ϕ. The extra overhead introduced by our association network is marginal, as compared to the standard GAN training cost. We did not notice per-iteration cost increase. Actually, we observed that with INTACT the whole model often converges using less epochs, leading to a faster training process than the standard multi-task GAN baseline. We set the learning rate to 1×10^{-4} for generator G and discriminator D, and 1×10^{-3} for the association module ϕ. The mini-batch size was 32. We set the loss hyper-parameters consistently in all the experiments as: $\lambda_g = 0.1$, $\lambda_c = 0.3$. In practice, we selected these parameters by balancing their loss value scales to avoid any dominating term in training.

4.4. Comparisons to State-of-the-Art Methods

We compared our INTACT with a wide range of state-of-the-art re-id methods, including (1) Conventional person re-id models: CamStyle [51] and FD-GAN [12]; (2) Super-resolution based models: SLD^{2}L [17], SING [16], CSR-GAN [40]; (3) Resolution-invariant representation learning based models: JUDEA [25], SDF [39], RAIN [7]; and (4) A hybrid method CAD [26] that combines SR and resolution-invariant representation learning.

The results comparisons are shown in Table 1. We have the following observations:

(1) INTACT achieves the state-of-the-art performance on all the five datasets, consistently outperforming the best competitor [26] by up to 6% at Rank-1.

(2) Compared to the SR based cross-resolution person re-id methods (SLD^{2}L [17], SING [16], CSR-GAN [40]), INTACT achieves significant improvement, *e.g.* up to 15.1% Rank-1 performance boost. This validates that our model can effectively address the inferior compatibility issue between image SR and person re-id as suffered by these previous state-of-the-art methods.

(3) Compared to resolution-invariant representation learning models (JUDEA [25], SDF [39], RAIN [7]), INTACT achieves the best performance on both the small datasets (MLR-VIPeR and CAVIAR, which is generally very challenging for deep learning methods due to no sufficient training data), and the large dataset (MLR-CUHK03), often by a large margin. This suggests that image SR based methods provide more superior solutions.

(4) Compared to the best competitor [26] that exploits both image SR and resolution-invariant representation learning, INTACT remains a better method by only using image SR as the core strategy.

(5) The standard person re-id models (CamStyle [51] and FD-GAN [12]) suffer from significant performance drop on MLR person re-id datasets, as compared to their reported results on standard HR person re-id datasets. This shows that the resolution mismatch problem is typically ignored by most existing re-id methods.

4.5. Inter-Task Association Analysis

We adopted a multi-task learning framework as our base model, where the SR module serves as a preprocessing step to recover the essential details originally missing in LR images in order to more accurately match HR gallery images. We consider that the SR task inherently may be not compatible to the identity matching task, the reason why we introduced INTACT as an explicit regularisation for SR. Here, we conducted an experiment to examine the association between the two different tasks (image SR & person re-id) and its effect on the overall model performance.

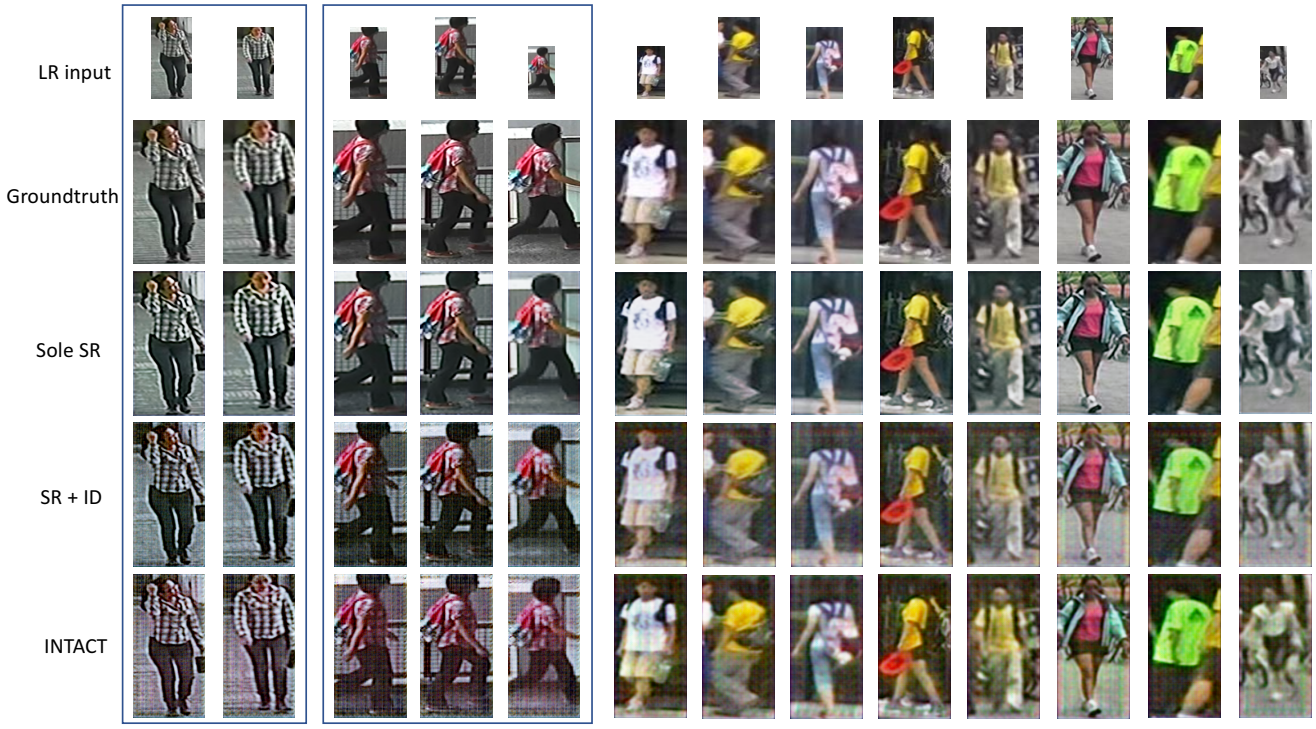

Figure 4. Qualitative examples of recovered test images from MLR-CUHK03 and MLR-Market1501.

Evaluation metrics We measured the pixel-wise SR quality of the recovered images by INTACT on the MLR-CUHK03 test set. We used the SSIM and PSNR metrics, together with Rank-1 as person re-id performance metric.

Competitors We compared INTACT with: (1) Sole SR: our method without person re-id constraint, (2) SR+ID: jointly learning image SR and person re-id, and (3) four state-of-the-art re-id models (CycleGAN [53], SING [16], CSR-GAN [40] and CAD [26]).

Results The performance comparisons in Table 2 show the following observations:
(1) The sole SR module of INTACT (supervised by MSE loss only) achieves the best pixel-wise SR performance, *i.e.* the highest PSNR and SSIM scores. This verifies the effectiveness of our SR generator.
(2) Although the sole SR model achieves the highest PSNR and SSIM performance, its resolved images yield the worst accuracy for cross-resolution person re-id. This indicates that as the low-level image quality metrics, both SSIM and PSNR are unsuited for evaluating high-level semantic recognition tasks such as person re-id in our case.
(3) The model in favour of SR performance, does not provide improvements on cross-resolution person re-id performance. This suggests that the SR supervision is not directly relevant to person re-id.
(4) The SR module, that does benefit the person re-id task, enhances only the identity matching related details whilst ignores other fine-grained details. This actually produces inferior pixel-level fidelity.

For qualitative examination, we showed several qualitative examples of recovered images in Fig. 4. Whilst our INTACT notably outperforms all the baselines in numerical evaluation, this performance difference is however seldomly reflected in the low-level image space. This implies that high-level semantic objective is less interpretive due to high functional complexity of deep network models.

Table 2. Comparison of super-resolution and cross-resolution person re-id performance on the MLR-CUHK03 test set.

Model	SSIM	PSNR	Rank1
CycleGAN [53]	0.55	14.1	62.1
SING [16]	0.65	18.1	67.7
CSR-GAN [40]	0.76	21.5	71.3
CAD [26]	0.73	20.2	82.1
Sole SR	**0.82**	**26.6**	23.0
SR + ID	0.77	23.3	82.7
INTACT	0.73	22.8	**86.4**

4.6. Ablation Study

Loss component analysis Our INTACT is jointly trained with image SR, person re-id and the inter-task association loss functions (cf. Eq. (7) & (8)). We examined their per-

formance effects on the MLR-Market-1501 dataset. Table 3 reports the ablation results. We observed that:
(1) With identity classification alone, the model achieves the poorest matching performance. This verifies that jointly learning the multi-task framework with the standard cascaded SR and person re-id model is unsatisfactory.
(2) After adding GAN loss, the model achieves slightly better performance. The plausible reason is that the adversarial loss helps to align the statistics of resolved images to the native HR data. However, the improvement is fairly marginal.
(3) Importantly, the proposed association loss brings a significant improvement, verifying the effectiveness of our regularisation scheme based on the idea of exploiting the underlying inter-task correlation.

Table 3. Evaluating INTACT's loss components on MLR-Market-1501. MSE: pixel-wise content loss, ID: identity classification loss (Eq. (3)), Association: our association loss (Eq. (7) & (8)).

Supervision	Rank1	Rank5	Rank10
MSE+ID	83.7	93.0	95.6
MSE+ID+GAN	84.7	93.9	96.1
MSE+ID+GAN+Association	**88.1**	**95.0**	**96.9**

Association design For the association learning between the discrimination and recognition representations in INTACT, we adopt a recognition-to-discriminator design. This is based on a hypothesis that the identity recognition representations learned from the HR images should contain the desired information of high resolution (that the real-fake HR discriminator tries to learn); And the real-fake HR discriminator representations, derived by a simple binary classification task, are relatively simpler.

We examined the effect of association design by additionally testing two more formulations: (i) Common space association (Fig. 5 (a)), and (ii) Discrimination-to-recognition association (Fig. 5 (b)) which is the inverse of the recognition-to-discriminator design (Fig. 5 (c)) we adopt in INTACT. Table 4 shows that different designs present fairly similar performances, and the recognition-to-discriminator is the best choice. This verifies the proposed association strategy.

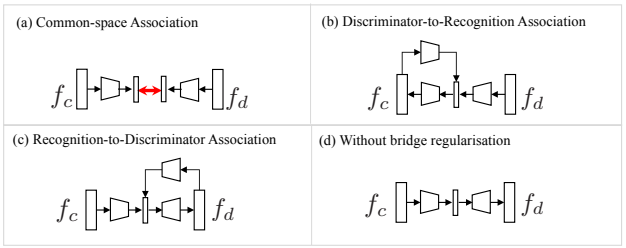

Figure 5. Schematics of different association designs.

Table 4. Evaluating the association design. R-to-D: from identity recognition representation to discriminator representation (used in INTACT); D-to-R: the inverse.

Association Space	Rank1	Rank5	Rank10
Common Space (a)	84.3	94.0	95.3
D-to-R (b)	83.4	93.5	95.0
R-to-D (c, ours)	**88.1**	**95.0**	**96.9**

Bridging constraint To facilitate the training of inter-task association between the discriminator and identity classification representations, we isolated an intermediate latent feature space f_e from ϕ to bridge the association target (Eq. (6) and (7)), implemented by an encoder-decoder structure. The result in Table 5 shows that the introduction of such a bridging constraint helps to better constrain the associative learning.

Table 5. Effect of the bridge constraint (Eq. (6)).

Bridge constraint	Rank1	Rank5	Rank10
W/O (Fig. 5 (d))	84.3	93.5	95.8
W (Fig. 5 (c))	**88.1**	**95.0**	**96.9**

5. Conclusion

In this work, we presented a novel deep learning regularisation, named *Inter-Task Association Critic* (INTACT), for solving the under-studied yet important cross-resolution person re-id problem. As a generic learning constraint, INTACT is designed specially for improving the training of existing multi-task (image SR and person re-id) models, by alleviating properly the difficulty of gradients backpropagation through two cascaded networks. During training, INTACT discovers the underlying association knowledge between image SR and person re-id by learning from the HR training data, and uses the self-discovered association information to further guide the learning behaviour of SR model alternatively. Thus the compatibility of SR with re-id matching can be maximised. This is built up on parameterising the inter-task association with a dedicated network. Extensive experimental results have demonstrated the performance superiority of our model over a wide variety of existing cross-resolution and standard person re-id methods on five challenging benchmarks. Component analysis of our method provides insights into the formulation of INTACT.

Acknowledgement

This work was supported by the Innovate UK Industrial Challenge Project on Developing and Commercialising Intelligent Video Analytics Solutions for Public Safety (98111-571149), the Alan Turing Institute Turing Fellowship, and Vision Semantics Limited.

References

[1] Song Bai, Peng Tang, Philip HS Torr, and Longin Jan Latecki. Re-ranking via metric fusion for object retrieval and person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 740–749, 2019. 2

[2] Sarath Chandar, Chinnadhurai Sankar, Eugene Vorontsov, Samira Ebrahimi Kahou, and Yoshua Bengio. Towards non-saturating recurrent units for modelling long-term dependencies. *arXiv preprint arXiv:1902.06704*, 2019. 2, 4

[3] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2109–2118, 2018. 2

[4] Jiawei Chen, Jonathan Wu, Janusz Konrad, and Prakash Ishwar. Semi-coupled two-stream fusion convnets for action recognition at extremely low resolutions. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 139–147. IEEE, 2017. 2

[5] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Deep association learning for unsupervised video person re-identification. *arXiv preprint arXiv:1808.07301*, 2018. 2

[6] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Instance-guided context rendering for cross-domain person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 232–242, 2019. 2

[7] Yun-Chun Chen, Yu-Jhe Li, Xiaofei Du, and Yu-Chiang Frank Wang. Learning resolution-invariant deep representations for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8215–8222, 2019. 1, 2, 6

[8] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE transactions on pattern analysis and machine intelligence*, 40(2):392–408, 2017. 2

[9] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. Low-resolution face recognition. In *Asian Conference on Computer Vision*, pages 605–621. Springer, 2018. 2

[10] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. Characteristic regularisation for super-resolving face images. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2435–2444, 2020. 2

[11] Qi Dong, Shaogang Gong, and Xiatian Zhu. Person search by text attribute query as zero-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3652–3661, 2019. 2

[12] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *Advances in Neural Information Processing Systems*, pages 1222–1233, 2018. 1, 6

[13] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy. *Person re-identification*. Springer, January 2014. 1

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[16] Jiening Jiao, Wei-Shi Zheng, Ancong Wu, Xiatian Zhu, and Shaogang Gong. Deep low-resolution person re-identification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 1, 2, 3, 5, 6, 7

[17] Xiao-Yuan Jing, Xiaoke Zhu, Fei Wu, Xinge You, Qinglong Liu, Dong Yue, Ruimin Hu, and Baowen Xu. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 695–704, 2015. 2, 6

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5

[19] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 3, 5

[20] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570, 2015. 2, 4

[21] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *Proceedings of the European conference on computer vision (ECCV)*, pages 737–753, 2018. 2

[22] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised tracklet person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2

[23] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2018. 2

[24] Wei Li, Xiatian Zhu, and Shaogang Gong. Scalable person re-identification by harmonious attention. *International Journal of Computer Vision*, pages 1–19, 2019. 2

[25] Xiang Li, Wei-Shi Zheng, Xiaojuan Wang, Tao Xiang, and Shaogang Gong. Multi-scale learning for low-resolution person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3765–3773, 2015. 2, 6

[26] Yu-Jhe Li, Yun-Chun Chen, Yen-Yu Lin, Xiaofei Du, and Yu-Chiang Frank Wang. Recover and identify: A generative dual model for cross-resolution person re-identification. *arXiv preprint arXiv:1908.06052*, 2019. 1, 2, 5, 6, 7

[27] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *Proceedings of the AAAI Confer-*

ence on Artificial Intelligence, volume 33, pages 8738–8745, 2019. 2

[28] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4099–4108, 2018. 2

[29] Ze Lu, Xudong Jiang, and Alex Kot. Deep coupled resnet for low-resolution face recognition. *IEEE Signal Processing Letters*, 25(4):526–530, 2018. 2

[30] Xiaolong Ma, Xiatian Zhu, Shaogang Gong, Xudong Xie, Jianming Hu, Kin-Man Lam, and Yisheng Zhong. Person re-identification by unsupervised video matching. *Pattern Recognition*, 65:197–210, 2017. 2

[31] Jingke Meng, Sheng Wu, and Wei-Shi Zheng. Weakly supervised person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 760–769, 2019. 2

[32] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 542–551, 2019. 2

[33] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *CVPR*, 2016. 2

[34] Michael S Ryoo, Kiyoon Kim, and Hyun Jong Yang. Extreme low resolution activity recognition with multi-siamese embedding learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[35] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 486–504, 2018. 2

[36] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 719–728, 2019. 2

[37] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2275–2284, 2018. 2

[38] Zhangyang Wang, Shiyu Chang, Yingzhen Yang, Ding Liu, and Thomas S Huang. Studying very low resolution recognition using deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4792–4800, 2016. 2

[39] Zheng Wang, Ruimin Hu, Yi Yu, Junjun Jiang, Chao Liang, and Jinqiao Wang. Scale-adaptive low-resolution person re-identification via learning a discriminating surface. In *IJCAI*, pages 2669–2675, 2016. 6

[40] Zheng Wang, Mang Ye, Fan Yang, Xiang Bai, and Shin'ichi Satoh. Cascaded sr-gan for scale-adaptive low resolution person re-identification. In *IJCAI*, pages 3891–3897, 2018. 1, 2, 6, 7

[41] Ancong Wu, Wei-Shi Zheng, and Jian-Huang Lai. Unsupervised person re-identification by camera-aware similarity consistency learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6922–6931, 2019. 2

[42] Guile Wu, Xiatian Zhu, and Shaogang Gong. Tracklet self-supervised learning for unsupervised person re-identification. In *AAAI Conference on Artificial Intelligence*, 2020. 2

[43] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1249–1258, 2016. 2

[44] Zhou Yin, Wei-Shi Zheng, Ancong Wu, Hong-Xing Yu, Hai Wan, Xiaowei Guo, Feiyue Huang, and Jianhuang Lai. Adversarial attribute-image person re-identification. *arXiv preprint arXiv:1712.01493*, 2017. 2

[45] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 994–1002, 2017. 2

[46] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jian-Huang Lai. Unsupervised person re-identification by soft multilabel learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2148–2157, 2019. 2

[47] Kaipeng Zhang, Zhanpeng Zhang, Chia-Wen Cheng, Winston H Hsu, Yu Qiao, Wei Liu, and Tong Zhang. Super-identity convolutional neural network for face hallucination. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 183–198, 2018. 2

[48] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):653–668, 2012. 2

[49] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Towards open-world person re-identification by one-shot group-based verification. *IEEE transactions on pattern analysis and machine intelligence*, 38(3):591–606, 2015. 2

[50] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 598–607, 2019. 2

[51] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2018. 6

[52] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *ICCV*, 2019. 2, 3, 4, 5

[53] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 7

[54] Xiangping Zhu, Xiatian Zhu, Minxian Li, Vittorio Murino, and Shaogang Gong. Intra-camera supervised person re-identification: A new benchmark. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2